Induction of Medical Expert System Rules based on Rough Sets and Resampling Methods

Shusaku Tsumoto and Hiroshi Tanaka
Department of Informational Medicine
Medical Research Institute,Tokyo Medical and Dental University
1-5-45 Yushima, Bunkyo-ku Tokyo 113 Japan
TEL: +81-3-3813-6111 (6159) FAX: +81-3-5684-3618
E-mail:{tsumoto, tanaka}@tmd.ac.jp

Abstract

Automated knowledge acquisition is an important research issue in improving the efficiency of medical expert systems. Rules for medical expert systems consists of two parts: one is a proposition part, which represent a if-then rule, and the other is probabilistic measures, which represents reliability of that rule. Therefore, acquisition of both knowledge is very important for application of machine learning methods to medical domains. Extending concepts of rough set theory to probabilistic domain, we introduce a new approach to knowledge acquisition, which induces probabilistic rules based on rough set theory(PRIMEROSE) and develop a program that extracts rules for an expert system from clinical database, using this method. The results show that the derived rules almost correspond to those of medical experts.

INTRODUCTION

One of the most important problems in rule induction methods is how to estimate the reliability of the induced results, which is a semantic part of knowledge to be induced from finite training samples. In order to estimate errors of induced results, resampling methods, such as cross-validation, the bootstrap method, have been introduced. However, while cross-validation method obtains better results in some domains, the bootstrap method calculates better estimation in other domains, and it is very difficult how to choose one of the two methods. In order to reduce these disadvantages further, we introduce the combination of repeated cross-validation method with the bootstrap method, both of which are studied as non-parametric error estimation methods or statistical model estimation ones in the community of statistics. The results show that this combination estimates the accuracy of the induced results correctly.

The paper is organized as follows: in section 2, we mention about probabilistic rules. Section 3 presents our new method, PRIMEROSE for induction of RHINOS-type rules. Section 4 gives experimental results. Finally, in section 5, we mention about Ziarko's related work, Variable Precision Rough Set Model.

RHINOS2 PROBABILISTIC RULES

Our approach is firstly motivated by automatic rule generation for RHINOS [5].RHINOS is an expert system which diagnoses the causes of headache or facial pain from manifestations. For the limitation of the space, in the following, we only discuss about the acquisition of inclusive rules,which are used for differential diagnosis. For further information,refer to [5].

Inclusive rule consists of several rules,which we call positive rules. The premises of positive rules are composed of a set of manifestations specific to a disease to be included for the candidates of disease diagnoses. If a patient satisfy one set of the manifestation of a inclusive rule, we suspect the corresponding disease with some probability. These rules are derived by asking the following questions in relation to each disease to the medical experts:*1.a set of manifestations by which we strongly suspect a corresponding disease. 2.the probability that a patient has the disease with this set of manifestations:SI(Satisfactory Index) 3.the ratio of the number the patients who satisfy the set of manifestations to that of all the patients having this disease:CI(Covering Index) 4.If sum of the derived CI(tCI) is equal to 1.0 then end. If not, goto 5. 5.For the patients suffering from this disease who do not satisfy all the collected set of manifestations,*

Due to difficulties with international mail delivery, this paper did not arrive in time to be placed in the proper order in the Proceedings. We apologize to the authors.

0195-4210/94/$5.00 © 1994 AMIA, Inc.

goto 1. An inclusive rule is described by the set of manifestations, and its satisfactory index. Note that SI and CI are given experimentally by medical experts. For example, let us consider an example of inclusive rules. Let us show an example of an inclusive rule of common migraine(CI=0.75) as follows:

If
history:paroxysmal, jolt headache:yes,
nature: throbbing or persistent,
prodrome:no, intermittent symptom:no,
persistent time: more than 6 hours, and
location: not eye,
Then we suspect common migraine (SI=0.9, CI=0.75).

Then SI=0.9 denotes that we can diagnose common migraine with the probability 0.9 when a patient satisfies the premise of this rule. And CI=0.75 suggests that this rule only covers 75 % of total samples which belong to a class of common migraine.

Formally, we can represent each positive rule as a tuple: $\langle d, R_i, SI_i(, CI_i)\rangle$, where d denotes its conclusion, and R_i denotes its premise. The inclusive rule is described as: $\langle\{\langle d, R_1, SI_1(, CI_1)\rangle, \cdots, \langle d, R_k, SI_k(, CI_k)\rangle\}, tCI\rangle$. where total CI(tCI) is defined as the sum of CI of each rule with the same conclusion:$\sum_i CI_i$.

ROUGH SETS AND PRIMEROSE

Rough set theory is developed and rigorously formulated by Pawlak[9]. This theory can be used to acquire certain sets of attributes which would contribute to class classification and can also evaluate how precisely these attributes are able to classify data.

For the limitation of space, we mention only how to extend the original rough set model to probabilistic domain, which we call PRIMEROSE(Probabilistic Rule Induction Method based on ROugh Sets). And we denote a set which supports an equivalence relation R_i by $[x]_{R_i}$ and we call it an *indiscernible set.* For example, if an equivalence relation R is supported by a set {1,2,3}, then $[x]_R$ is equal to {1,2,3} ($[x]_R = \{1, 2, 3\}$).

Definition of Probabilistic Rules

We extend the definition of consistent rules to probabilistic domain. For this purpose, we use the definition of inclusive rules which Matsumura et.al [5] introduce for the development of a medical expert system, RHINOS(Rule-based Headache and facial pain INformation Organizing System). This inclusive rule is formulated in terms of rough set theory as follows:

Definition 1 (Probabilistic Rules) *Let R_i be an equivalence relation and D denotes a set whose elements belong to one class and which is a subset of U. A probabilistic rule of D is defined as a tuple, $< D, R_i, SI(R_i, D), CI(R_i, D) >$ where R_i, SI, and CI are defined as follows.*

R_i is a conditional part of a class D and defined as:

$$R_i \quad s.t. \quad [x]_{R_i} \bigcap D \neq \phi$$

SI and CI are defined as:

$$SI(R_i,D) = \frac{card\ \{([x]_{R_i} \bigcap D) \bigcup ([x]^c_{R_i} \bigcap D^c)\}}{card\ \{[x]_{R_i} \bigcup [x]^c_{R_i}\}}$$

$$CI(R_i,D) = \frac{card\ \{([x]_{R_i} \bigcap D) \bigcup ([x]^c_{R_i} \bigcap D^c)\}}{card\ \{D \bigcup D^c\}}$$

where D^c or $[x]^c_{R_i}$ consists of unobserved future cases of a class D or those which satisfies R_i,respectively. □

In the above definition, *unobserved future cases* means all possible future cases. So we consider an infinite size of cases, which is called *total population* in the community of statistics.

And SI(Satisfactory Index) denotes the probability that a patient has the disease with this set of manifestations, and CI(Covering Index) denotes the ratio of the number the patients who satisfy the set of manifestations to that of all the patients having this disease. Note that SI(R_i,D) is equivalent to the accuracy of R_i.

A total rule of D is given by $R = \bigvee_i R_i$, and then total CI(tCI) and total SI(tCI) is defined as: tCI(R,D) = CI($\bigvee_i R_i$,D), and tCI(R,D) = SI($\bigvee_i R_i$,D) respectively.

Since the above formulae include unobserved cases, we are forced to estimate these measures from the training samples. For this purpose, we introduction cross-validation and the Bootstrap method to generate "pseudo-unobserved" cases from these samples as shown in the next subsection.

Cross-Validation and the Bootstrap

Cross-validation method for error estimation is performed as following: first, the whole training samples $\mathcal{L}$ are split into V blocks: $\{\mathcal{L}_1, \mathcal{L}_2, \cdots, \mathcal{L}_V\}$. Second, repeat for V times the procedure in which we induce rules from the training samples $\mathcal{L} - \mathcal{L}_i (i = 1, \cdots, V)$ and examine the error rate err_i of the rules using $\mathcal{L}_i$ as test samples. Finally, we derive the whole error rate err by averaging err_i over i, that is, $err = \sum_{i=1}^{V} err_i/V$ (this method is called V-fold cross-validation). Therefore we can use this method for estimation of CI and SI by replacing the calculation of err by that of CI and SI, and by regarding test samples as unobserved cases.

On the other hand, the Bootstrap methods is executed as follows: first, we create empirical probabilistic distribution(F_n) from the original training samples. Second, we use the Monte-Carlo methods and randomly take the training samples by using F_n. Third, rules are induced by using new training samples. Finally, these results are tested by the original training samples and statistical measures, such as error rate are calculated. We iterate these four steps for finite times. Empirically, it is shown that about 200 times repetition is sufficient for estimation.

Interestingly, Efron[3] shows that estimators by 2-fold cross-validation are asymptotically equal to predictive estimators for completely new pattern of data, and that Bootstrap estimators are asymptotically equal to maximum likelihood estimators and are a little overfitting to training samples. Hence, we can use the former estimators as the lower bound of SI and CI, and the latter as the upper bound of SI and CI.

Furthermore, in order to reduce the high variance of estimators by cross-validation, we introduce repeated cross-validation method,which is firstly introduced by Walker[12]. In this method, cross-validation methods are executed repeatedly(safely, 100 times), and estimates are averaged over all the trials. In summary, since our strategy is to avoid the overestimation and the high variabilities, we adopt combination of repeated 2-fold cross-validation and the Bootstrap method in this paper.

Cluster-based Reduction of Knowledge

Reduction technique removes dependent variables from rules. This dependence is originated from algebraic dependence, that is , if $f(a_1, a_2, \cdots, a_n, a_{n+1}) = f(a_1, a_2, \cdots, a_n) = 0$ then a_{n+1} is dependent on $a_1, a_2, \cdots, a_n$. Hence, intuitionally, if the removal of one variable does not change the former consistent classification, we can remove this variable. In PRIMEROSE, we extend the concept of reduction to probabilistic domain: we delete an attribute when the deletion does not make apparent SI change. For example, if one rule support one class with some probability and other classes with some probabilities, we minimize its conditionals by the cluster-based reduction: that is, if the removal of one attribute does not change the above probabilities, we can remove this attribute.

This process means that we fix the probabilistic nature of the induced rules and is very effective when databases include inconsistent samples. This method is very similar to VPRS model introduced by Ziarko[13].

On the other hand, in the original Pawlak's models inconsistent parts is ignored and only reduction of the consistent parts is executed. For precise information, please refer to [9, 13, 11].

Algorithm for PRIMEROSE

Algorithms for rule induction can be derived by embedding rough set theory concept into the algorithms discussed in Section 2. An algorithm for induction of inclusive rules is described as follows:

1)Using all attributes, calculate all equivalent relation $\{R_i\}$ which covers all of the training samples, that is, calculate $\{R_i | \bigcup [x]_{R_i} = U\}$.
2)For each class D_j, collect all the equivalent relation R_i such that $[x]_{R_i} \cap D_j \neq \phi$. For each combination, calculate its possible region.
3)Calculate SI(R_i,D_j).
4)Apply probabilistic reduction of knowledge to each relation R_i until SI is changed(Minimize the components of each relation). If several candidates of minimization are derived, connect each with disjunction.
5)Collect all the rules, perform the cross-validation method and the bootstrap method to estimate utCI for each D_j.

EXPERIMENTAL RESULTS

We apply PRIMEROSE to headache(RHINOS's domain), meningitis, and cerebrovascular diseases, whose precise information are given in Table 2 and 3. These data are incomplete, and include many inconsistencies.

The experiments are performed by the following three procedures. First, we randomly splits these samples into pseudo-training samples and pseudo-test samples. Second, by using the pseudo-training samples, PRIMEROSE induces rules and the statistical measures. Third, the induced results are tested by the pseudo-test samples. We perform these procedures for 100 times and average each accuracy and the estimators for accuracy over 100

Table 1: Information of Database

Domain	Samples	Classes	Attributes
headache	121	10	20
meningitis	99	3	25
CVD	137	6	27

Table 2: Experimental Results (Comparison)

Domain	Method	Accuracy
headache	CART	62.8%
	AQ15	61.2%
	PRIMEROSE	74.4%
meningitis	CART	60.6%
	AQ15	67.7%
	PRIMEROSE	74.7%
CVD	CART	65.7%
	AQ15	73.0%
	PRIMEROSE	81.7%

trials. We compare PRIMEROSE with AQ15[6] and CART[1].

Experimental results are shown in Table 2 and 3. In Table 3, the first column shows estimators tested by the pseudo-test samples, as shown above. The second and third column denotes cross-validation estimator and the bootstrap estimator, respectively.

These results suggest that PRIMEROSE performs a little better than the other two methods and that the estimation of accuracy performs very well.

RELATED WORKS

Comparison with AQ15

AQ is an inductive learning method based on incremental STAR algorithm developed by Michalski [6]. This algorithm selects one seed from positive examples and starts from one "selector"(attribute). It adds selectors incrementally until the "complexes" (conjunction of attributes) explain only positive examples. Since many complexes can satisfy these positive examples, according to a flexible extra-logical criterion,AQ finds the most preferred one.

It would be surprising that the complexes supported only by positive examples corresponds to the positive region. That is, the rules induced by AQ is equivalent to consistent rules introduced by Pawlak[9]. However, as shown in [9], the ordinary rule induction by rough set theory is different from AQ in strategy;Pawlak's method starts from description by total attributes, and then performs reduction to get minimal reducts,that is, rules are derived in a top-down manner. On the contrary, AQ induces in a bottom-up manner. While these approaches are different in strategies, they are often equivalent because of logical consistency, and this difference suggests that when we need the large number of attributes to describe rules, induction based on rough set theory is faster.

One of the important problem of the AQ method is that it does not work well in probabilistic domain [6]. This problem is also explained by matroid theory: inconsistent data do not satisfy the condition of independence, so we cannot derive a basis of matroid in probabilistic domain using the proposed definition, which is the same problem as the Pawlak's method,as discussed in Section 4. Hence it is necessary to change the definition of independence to solve those problems.

As discussed earlier,in PRIMEROSE, we adopt cluster membership as the condition of independence, instead of using class membership. Restricting the probabilistic nature, we can use almost the same algorithm as class-consistency based reduction. Then we estimate the probabilistic nature of the derived rules using some resampling plans, such as cross-validation method in this paper. This is one kind of solution to the above problems, and the similar approach can also solve the disadvantage of AQ.

Table 3: Experimental Results (Estimation)

Domain	Test	CV	BS
headache	74.4%	58.7%	91.6%
meningitis	74.7%	59.6%	88.3%
CVD	81.7%	70.1%	87.5%

Comparison with CART and ID3

Induction of decision trees, such as CART[1] and ID3[10] is another inductive learning method based on the ordering of variables using information entropy measure or other similar measures. This method splits training samples into smaller ones in a top-down manner until it cannot split the samples, and then prunes the overfitting leaves.

There are many discussions about the problems of this approach[7, 8]. Two of the important problems are about high computational costs of pruning and structural instability. As shown in [4], constructing optimal binary decision trees is NP-complete. In this context, this means that it is difficult to determine which leaves should be pruned. CART uses the combination of cross validation method and minimal cost complexity. The dificulty is to calculate the complexity because we should choose the pruned leaves.

PRIMEROSE method also has the similar problems since reduction technique corresponds to pruning. While reduction technique examines the dependencies of attributes, pruning techniques are mainly based on the trade-off between accuracy and structural complexity.

Note that reduction technique only uses topological characteristics of the training samples. And dependencies and independencies of the attributes are important factors, since dependent attributes will not change accuracy of the induced rules. Moreover,as shown in [2], if the attributes are independent and quantized to k levels, there is no peaking phenomenon of accuracy in the Bayesian context, as discussed in the test-sample accuracy of decision trees.

Hence extracting independent variables is very important in probabilistic domain. These facts suggests that when the attributes are the mixture of dependent and independent ones, PRIMEROSE performs much better.On the other hand, when almost all of the attributes are independent, PRIMEROSE is much worse since we cannot use information about dependencies.

CONCLUSION

We introduce a new approach to knowledge acquisition, PRIMEROSE, and develop an program based on this method to extract rules for an expert system from clinical database. It is applied to three medical domains. The results show that the derived rules performs a little better than CART and AQ15 and that the estimation of statistical measures performs well.

Acknowledgements

The authors would like to thank Prof.Ziarko, Prof.Grzymala-Busse and Prof.Nitin Indurkhya for giving me some comments on the manuscript. This research is supported by Grants-in-Aid for Scientific Research No.04229105 from the Ministry of Education, Science and Culture, Japan.

References

[1] Breiman,L.,Freidman,J.,Olshen,R.,and Stone,C. *Classification And Regression Trees.* Belmont,CA:Wadsworth International Group, 1984.

[2] Chandrasekaran,B.and Jain,A.K. Quantization complexity and independent measurements *IEEE Trans.Comput.*,**23**,102-106,1974.

[3] Efron B. *The Introduction to the Bootstrap* Chapman-Hall, 1994. Pennsylvania: CBMS-NSF, 1982.

[4] Hyafil,L and Rivest,R.L. Constructing Optimal Binary Decision Trees is NP-complete. *Information Processing Letters*,1976.

[5] Matsumura,Y, et al. Consultation system for diagnoses of headache and facial pain: RHINOS,*Medical Informatics*,**11**,145-157,1986.

[6] Michalski,R.S.,et al. The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains, *Proc. of AAAI-86*, 1041-1045,Morgan Kaufmann,1986.

[7] Mingers,J. An Empirical Comparison of Selection Measures for Decision Tree Induction. *Machine Learning*,**3**,319-342, 1989.

[8] Mingers,J. An Empirical Comparison of Pruning Methods for Decision Tree Induction. *Machine Learning*,**4**,227-243, 1989.

[9] Pawlak,Z *Rough Sets*,Kluwer Academic Publishers, 1991.

[10] Quinlan, J.R. Induction of decision trees, *Machine Learning*, **1**, 81-106, 1986.

[11] Tsumoto,S. and Tanaka,H. PRIMEROSE: Probabilistic Rule Induction based on Rough Sets and Resampling MEthods, *Proc. of RSKD'93*, 1993.

[12] Walker,M.G. and Olshen,R.A. Probability Estimation for Biomedical Classification Problems. *Proc. of SCAMC-92*,McGrawHill,1992.

[13] Ziarko,W. Variable Precision Rough Set Model, *Journal of Computer and System Sciences*,**46**,39-59,1993.